DOCUMENT RESUME

ED 079 401 TM 003 005

AUTHOR Hollmann, Thomas D.
TITLE Differential Validity: A Problem with Tests or Criteria?
PUB DATE May 73
NOTE 12p.; Paper presented at a meeting of the Midwestern Psychological Association, May, 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Correlation; Evaluation Criteria; *Performance Criteria; Predictive Measurement; Predictor Variables; Racial Discrimination; *Rating Scales; *Test Bias; Testing Problems; *Test Interpretation; *Test Validity

ABSTRACT

The evidence used in condemning a test as racially biased is usually a validity coefficient for one racial group that is significantly different from that of another racial group. However, both variables in the calculation of a validity coefficient should be examined to determine where the bias lies. A study was conducted to investigate the construct validity of a set of predictors and a set of criterion scales separately for blacks and whites. Data were collected during a project to validate a test battery and a followup study of the effectiveness of the resultant selection procedures. Ss totaled 70 blacks and 104 whites. Tests used were the Adaptability Test, the Spelling scale, word meaning scale, checking scale and copying scale of the Purdue Clerical Adaptability Test and the ten scales of the Guilford-Zimmerman Temperament Survey. The criterion scales consisted of 5 specially constructed, behaviorally anchored rating scales measuring accuracy, information, attitude, initiative, and knowledge of procedures. Eleven cases of differential validity occurred. The pattern of test-criterion relationships was obviously different for the two groups. The factor patterns for the 15 predictors were for the most part similar for blacks and whites. The analysis of the factor structure of the predictors suggested that they measure the same or similar constructs within the two groups. Results indicate that the criterion scales, more than the predictors, contributed to the differential validity and single-group validity. (KM)

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 079401

TM 003 005

DIFFERENTIAL VALIDITY

A PROBLEM WITH TESTS OR CRITERIA?

Thomas D. Hollmann
Wayne State University

Presented at Midwestern
Psychological Association
May, 1973

The phrase "unfair test discrimination" suggests that this much discussed problem can be attributed to the tests used to predict later job performance. The evidence used to demonstrate the phenomenon consists of showing that the validity coefficient for one racial group is significantly different from that of another racial group. The implication is that the tests are measuring different constructs within different racial subgroups. A common extension of this line of reasoning is to argue against tests and their use, except in those cases where the validity of the test is the same for all racial groups involved.

There is, however, another factor which must be investigated before tests are branded as the culprits. It takes two variables to permit the calculation of a validity coefficient, and both of the variables must be investigated in terms of their contribution to differential validity. If psychological tests are capable of tapping different constructs in different racial groups, then so might the common types of subjective criterion rating scales.

Boehm, in her excellent review of studies showing differential validity and single group validity, suggests that the criterion may be responsible when job performance is measured by supervisory ratings. This is supported by Campbell, Pike, and Flaugher (1969) and by Bass and Turner (1973), both finding differences in criterion ratings which were related to ethnic group membership.

This study represents an attempt to investigate the construct validity of both a set of predictors and a set of criterion scales, separately for Blacks and Whites. The data were collected during a

project to validate a test battery and a follow-up study of the effectiveness of the resultant selection procedures. This secondary analysis was performed to help understand obtained examples of differential and single group validity.

The original validation study was conducted with a sample of 88 clerical employees in an insurance company; 30 Blacks and 58 Whites. One year after the new selection procedures were instituted, criterion data were collected for all clerical employees hired during the first nine months of that period. This group consisted of 40 Blacks and 46 Whites. Since this analysis focused on the meaning of the measures rather than on prediction, the two groups of subjects were combined for a grand total of 70 Blacks and 104 Whites.

The validation study was a modified concurrent design. The sample consisted of employees who had been tested before being hired; however, the test were not used in any systematic way in the hiring decision. Tests used were the Adaptability Test, the Spelling scale, Word Meaning scale, Checking scale and Copying scale of the Purdue Clerical Adaptability Test, and the ten scales of the Guilford-Zimmerman Temperament Survey. The criterion consisted of five behaviorally anchored rating scales constructed specifically for this study. The scales measured dimensions defined as Accuracy, Information, Attitude, Initiative, and Knowledge of Procedures.

There were no statistically significant differences between the two racial groups on any of the 15 predictor scales or on any of the five criterion scales in the initial study. However, the pattern of intercorrelations did reveal many between group differences. While there were some significant predictor-criterion correlations within each group, there was only one predictor-criterion pair

which was significant for both groups. Step-wise multiple regression was used to develop prediction equations for each group. Significant multiple R's were obtained for each criterion scale for each sub-group, although the equations were markedly different. It was assumed at the time that the differences in the patterns of intercorrelations were due to the differential meanings of the tests (test bias, if you must) for the two racial groups.

No cross-validation was performed on the original data; rather, since the subjects had been on the job for a while, it was decided that a follow-up study should be done to check the validity of the predictions to the criterion scales using a sample of new employees. This was done, and hence the additional subjects for the secondary analysis.

At this point I became interested in exploring the reasons underlying the differences in the validities of the various predictors. The impetus came from observed similarities in the patterns of intercorrelations among the predictors for the two groups. In terms of the construct validity paradigm, these similarities were not consistant with the belief that the predictors were tapping different constructs in each of the racial subgroups.

One way of determining the construct validity of a test is to factor analyze the test along with a battery of tests of known meaning. The meaning of the new test is then derived from its relationships with the constructs measured by the known tests. It follows that if the factor structure of a test battery is the same for two populations, the tests are measuring the same constructs in both.

The methodology used in this study followed this logic. First, the 15 predictor scales were factor analyzed separately for the two groups, and the factor structures compared. Then, again within each group separately, each criterion scale was factored with the 15 predictors. This allowed a between group comparison for each criterion scale in terms of how each was related to the dimensions being measured by the predictors.

RESULTS

For the total sample there was just one instance of a predictor-criterion pair which was significant in the same direction for both groups. There were 11 cases of differential validity; two of these were instances of the validity coefficients being significant for both groups and significantly different from each other (and in this case, significantly positive for one group and significantly negative for the other). There were also 13 cases of single group validity, three significant just for Blacks and ten significant just for Whites. The pattern of test-criterion relationships is obviously very different for the two groups.

The factor patterns for the 15 predictors are for the most part similar for Blacks and Whites (see Table 1.), the only striking difference being the greater differentiation on the ability tests for Whites. This not unexpected finding is the result of relatively subtle differences in the intercorrelation matrix. The whole pattern of intercorrelations among the five ability measures is higher for Whites than for Blacks. However, the correlations between the

Adaptability test and the Word Scale and between the Checking scale and the Copy scale are so much higher that they result in two factors being defined instead of one.

This analysis of the factor structure of the predictors suggests that they are in fact measuring the same or similar constructs within each of the racially defined groups. If this were not so, the probability of such similar factor structures would be low, indeed. There is no need to define or name the constructs being measured by the tests at this time. Suffice it to say that whatever they are, they are the same for both groups.

The criterion variables involved in the greatest number of instances of differential validity were Attitude and Initiative. When the Attitude criterion measurements are included in a factor analysis with the predictors (see Table 2.) we find that it is positively correlated with Factor II for Whites and negatively correlated with Factor II for Blacks. The results from the analysis of the Initiative scale with the predictors (see Table 3.) are not so clear-cut but do again demonstrate the problem. Initiative is not related to any of the common factors for Whites, while it loads on a previously undetected factor for Blacks.

The criterion variables of Procedures and Information were most often involved in single group validity situations. The Knowledge of Procedures scale (see Table 4.) was related to Factor III for Whites, but to a previously undetected factor for Blacks. From the correlation matrix we find that for Blacks, the correlations between this criterion scale and the Adaptability Test and the Word scale which define Factor III for Whites are -.09 and -.06 respectively.

The Information scale (see Table 5.) had the most disruptive effect on the factor structure when added to the analysis. It is related to the Word Meaning test scale for Blacks and to Factor VI, a personality dimension, for Whites.

The accuracy scale was the least well predicted to of the lot. Again, factoring it in conjunction with the predictor scales shows that it is related to different predictor dimensions (see Table 6.). For Whites it is related to Factor IV, while for Blacks it is related to a weakly defined personality dimension. Also, for Blacks, the Accuracy scale correlated -.05 and .04 with the Adaptability Test and the Word Meaning scale which define Factor IV for Whites.

Discussion and Conclusions

These results indicate that the criterion scales, more than the predictors, are contributing to the differential validity and single group validity in this study. The factor structures for the predictors are too similar to warrant the conclusion that the tests are measuring different constructs in the different racial groups. It is also worth noting that the commonalities for each of the criterion scales were, with two exceptions, almost identical to the squared multiple R's obtained for the prediction equations. This means that the majority of the valid variance is common variance in the current analysis and that there are no minor factors, undetected here, which could be different for the racial groups and therefore account for the differential validity.

Exactly what is happening is open to interpretation and will

require more study. There appears to be a general trend in the data indicating that the performance ratings of Blacks are related to personality dimensions although not usually the major personality dimensions defined by the predictor scales in this study. This is not so evident for Whites, especially when we note that the criterion scales of Knowledge of Procedures and Accuracy, the two real ability scales, are related to ability measures in the predictor domain.

This study also raises an interesting point. By all objective indices, the validation study which yielded these data resulted in a selection procedure which complies with the EEOC guidelines. There were no racial differences for mean scores on predictors or criterion scales. Separate regression equations were developed for each racial group. Use of the equations did not have an adverse impact, in fact new hires were running very close to 50-50 in an urban area which is almost 50% Black. Yet we find racial differences in the type of person who is considered an effective performer on the job. Therefore, while such procedures may pass muster based on current standards, the evidence indicates another example of equal but different treatment.

One can only speculate about the long run effects. If such procedures are common it could have an effect on the evolution of personality structure in two segments of the population by differentially rewarding personality types. More pragmatically, such practices may compound the problems in eliminating racial discrimination in promotion decisions. Over the years the organization will amass two different populations distinguishable on the dimension of race. If the abilities and personality characteristics necessary for

success at the higher level position are different than at the lower level, then it is very likely that the base rate for success will be different for the two populations. The organization will then have to promote persons not likely to succeed or have their promotion procedures subjected to attack on the basis of adverse impact. Therefore, what looks like a solution to unfair discrimination at one level of personnel procedures may lead to greater problems at another level.

Finally, the evidence reported here suggests that it may be less important to worry about discrimination based on test usage and more important to concentrate on fairness in performance ratings.

References

Bass, A.R. and Turner, J.N. Ethnic group differences in relationships among criteria of job performance. Journal of Applied Psychology. 1973, 57, 101-109.

Boehm, V.R. Negro white differences in validity of employment and training selection procedures: summary of research evidence. Journal of Applied Psychology. 1972, 56, 33-39.

Campbell, J.T., Pike, L.W., and Flaugher, R.L. Prediction of job performance for Negro and White medical technicians-A regression analysis of potential test bias: Predicting job knowledge scores from an aptitude-battery. (Educational Testing Service Rep. PR-69-6) Princeton, N.J.: Educational Testing Service, 1969.

T. D. Hollmann
MPA May, 1973

Table 1. Predictor factor structure for Blacks and Whites.

	Blacks				Whites					
	I	II	III	IV	I	II	III	IV	V	VI
Adaptability				48				80		
Word				60				69		
Checking				60					55	
Copy				52					65	
Spelling				65						-34
Gen. Activity	44				68					
Ascendancy	69				71					40
Sociability	88				64					
Thoughtfulness	34				34					
Emotional Stab.		60				53	33			44
Objectivity		93				90				
Friendliness		33	76			53	46			
Personal Relat.		36	43			70				
Restraint			74				50			
Masculinity										53
% Total Var.	11.6	11.2	9.8	11.7	10.6	14	5.7	9.5	7.1	6.7

Table 2. Factor structures with criterion scale: ATTITUDE

	Blacks				Whites					
	I	II	III	IV	I	II	III	IV	V	VI
Attitude		-33				27				
Adaptability				48				86		
Word				72				64		-34
Checking				52					66	
Copy				45					60	
Spelling				61						-51
Gen. Activity	43				65					
Ascendancy	73				74					
Sociability	84				65					
Thoughtfulness	36				36		-44			
Emotional Stab.		67				64				41
Objectivity		80				84				
Friendliness		30	76			30	76			
Personal Relat.		37	43			72				
Restraint			76							
Masculinity										33

T. D. Hollmann
MPA May, 1973

Table 3. Factor structure with criterion scale: INITIATIVE

	Blacks					Whites					
	I	II	III	IV	V	I	II	III	IV	V	VI
Initiative					-62						
Adaptability				47					78		32
Word				61					70		
Checking				67	38					51	
Copy				52						66	
Spelling				63					31		-31
Gen. Activity	40				30	69					
Ascendancy	77					74					35
Sociability	82					63					
Thoughtfulness	35					35					
Emotional Stab.		65			37		53	30			41
Objectivity		84					90				
Friendliness		34	34				55	48			
Personal Relat.		37	43				70				
Restraint			74					49			
Masculinity											54

Table 4. Factor structure with criterion scale: KNOWLEDGE OF PROCEDURES

	Blacks					Whites					
	I	II	III	IV	V	I	II	III	IV	V	VI
Knowledge					66				39		
Adaptability				48					95		
Word				60					48		-47
Checking				63						58	
Copy				53						61	
Spelling				64							-53
Gen. Activity	39				-34	64					35
Ascendancy	77					77		-37			
Sociability	82					61					
Thoughtfulness	36					38					
Emotional Stab.		59					64				44
Objectivity		91					92				
Friendliness		31	77				45	59			
Personal Relat.		35	43				69				
Restraint			74					34			
Masculinity											31

FILMED FROM BEST AVAILABLE COPY

T. D. Hollmann
MPA May, 1973

Table 5. Factor structure with criterion scale: INFORMATION

	Blacks					Whites					
	I	II	III	IV	V	I	II	III	IV	V	VI
Information					34						42
Adaptability				36					73		37
Word				40	57				70		
Checking				82						52	
Copy				53						67	
Spelling				52					37		
Gen. Activity	44					67					
Ascendancy	75					78					31
Sociability	83					62					
Thoughtfulness	36					38					
Emotional Stab.		70					52	31			38
Objectivity		82					86				
Friendliness		36	73				58	37			
Personal Relat.		39	42				70				
Restraint			76					52			
Masculinity											59

Table 6. Factor structure with criterion scale: ACCURACY

	Blacks					Whites					
	I	II	III	IV	V	I	II	III	IV	V	VI
Accuracy					63				29		
Adaptability				45					90		
Word		32		56					58		39
Checking				71						60	
Copy				54						62	
Spelling				63							50
Gen. Activity	44					64					
Ascendancy	73					75		-38			
Sociability	86					61					
Thoughtfulness	33				31	40					
Emotional Stab.		57			-31		59				-47
Objectivity		81					92				
Friendliness		31	77				49	46			
Personal Relat.		36	43				70				
Restraint			74					51			
Masculinity											-40